

Periodic Transcriptional Organization of the *E. coli* Genome

François Képès^{1,2*}

¹Atelier de Génomique
Cognitive, CNRS
UMR8071/genopole®, 523
Terrasses de l'Agora, 91000
Evry, France

²Epigenomics Project
genopole®, Evry, France

The organization of transcription within the prokaryotic nucleoid may be expected to both depend on and determine the structure of the chromosome. Indeed, immunofluorescence localization of transcriptional regulators has revealed foci in actively transcribing *Escherichia coli* cells. Furthermore, structural and biochemical approaches suggest that there are ~50 independent loop domains per genome. Here I show that in four *E. coli* strains, genes that are controlled by a sequence-specific transcriptional regulator tend to locate next to the gene encoding this regulator, or at regular distances that are multiples of 1/50th of the chromosome length. This periodicity is consistent with a solenoidal epigenetic organization of the chromosome, which would gather into foci the interacting partners; the regulator molecules and their DNA binding sites. Binding at genuine regulatory sites on DNA would thus be optimized by co-transcriptionally translating regulators in their vicinity.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: replication; transcription; *E. coli*; DNA functional organization; DNA-binding proteins

*Corresponding author

Introduction

In prokaryotes, the predominant level of control of gene expression involves regulator proteins that bind to short DNA sequence motifs in the *cis*-regulatory region of a target operon or gene, thus activating or repressing its transcription by RNA polymerase. Multivalent regulators bind to closely spaced sites in the regulatory region of a target, and this close spacing is in part responsible for increasing the local concentration of the regulators such that binding to one site facilitates binding to a second site.^{1,2} Because transcription and translation occur in the same compartment, one additional strategy to achieve this local concentration effect could be to position the regulator-encoding cistron close to its targets along the linear structure of the DNA: “1-D clustering”. The regulator is consequently produced near its targets in the frequent case where its synthesis initiates from messengers that are still tethered to DNA by ongoing transcription.³ Self-regulation is obviously

one case of 1-D clustering. It is frequently observed,^{4–6} but an open question is whether it accounts for all the cases of 1-D clustering.

Beyond the limits of linear contiguity, another means to increase local concentration is through 3-D spatial proximity: “3-D clustering”. If the numerous targets of a potent regulator were scattered throughout the chromosome, they could dynamically self-organize into a focus around the regulator-encoding cistron. Recent immunofluorescence studies indicate that transcription foci may occur in bacteria. In *Escherichia coli* cells, some sequence-specific DNA-binding proteins show a punctate pattern of fluorescence.⁷ In living *Bacillus subtilis* cells, green fluorescent protein (GFP)-tagged RNA polymerases are clustered in two to four discrete foci that comprise the most active rRNA encoding operons.⁸ As loci that are distant along the chromosome are brought together by 3-D clustering, long-range DNA looping is predicted. Indeed, topological studies suggest that the chromosome in growing *E. coli* cells is segregated into 40 or more independent domains of supercoiling.^{9,10} Remarkably, the number of independent loops or foci depends on cellular growth rate, suggesting that these specific topologies reflect a dynamic process.^{8,11,12}

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2004.05.039

E-mail address of the corresponding author:
francois.kepes@genopole.cnrs.fr

Sufficient data are now available about the transcriptional interactions in *E. coli*^{5,6} to explore the occurrence of 3-D and 1-D regulator/target clustering, its consequence on target/target colocalization, and its relationship with transcription foci.

Periodicity of distances between regulator-encoding cistron and targets

For each of 116 regulators, the dataset used here provides a list of targets mined from the literature.⁶ Regularities were sought in the distances between the translation startpoints of the regulator-encoding cistron and of each of the targets of this regulator (for operon targets, only the first cistron was considered). This process was repeated for each of the 116 regulators. The distribution of these 555 distances was then cumulated for all regulators (Figure 1a). Because the *E. coli* chromosome is circular, these distances are less than 2320 kbp, i.e. one half of the total circumference. The distribution has a peak for the shortest distances, indicative of 1-D clustering: among 128 occurrences, 58 cases (45%) are accounted for by self-regulation, and 70 cases correspond to regulation of a neighboring target that does not comprise the regulator-encoding

cistron. This result suggests that self-regulation is driven in part by the mechanistic principle that more generally underlies 1-D clustering. It remains possible that self-regulation reflects additional constraints such as response delay,¹³ biosynthetic cost or dynamical stability.¹⁴

However, the observed distribution differs significantly from a random distribution for longer distances too, with a *P*-value around 2% by the Kolmogorov–Smirnov test (see Supplementary Table 1 online). The regulator-encoding cistrons tend to be periodically spaced from their targets (Figure 1a). Thus, the remainder of the division of the regulator/target distance by the period should be smaller for the correct period than for any other one. If furthermore the correct period were identical for all regulator/target pairs, the remainders averaged over all pairs should be minimal. This principle was used to determine the period. Starting at 1 kbp, the period was incremented in 1 kbp steps and used as the divisor to compute the average remainders (see lower curve in Supplementary Figure 1 online). The first minimum was reached for a period around 92–96 kbp. This estimate of the period was manually refined to 92.8 kbp by maximizing the number of periodic

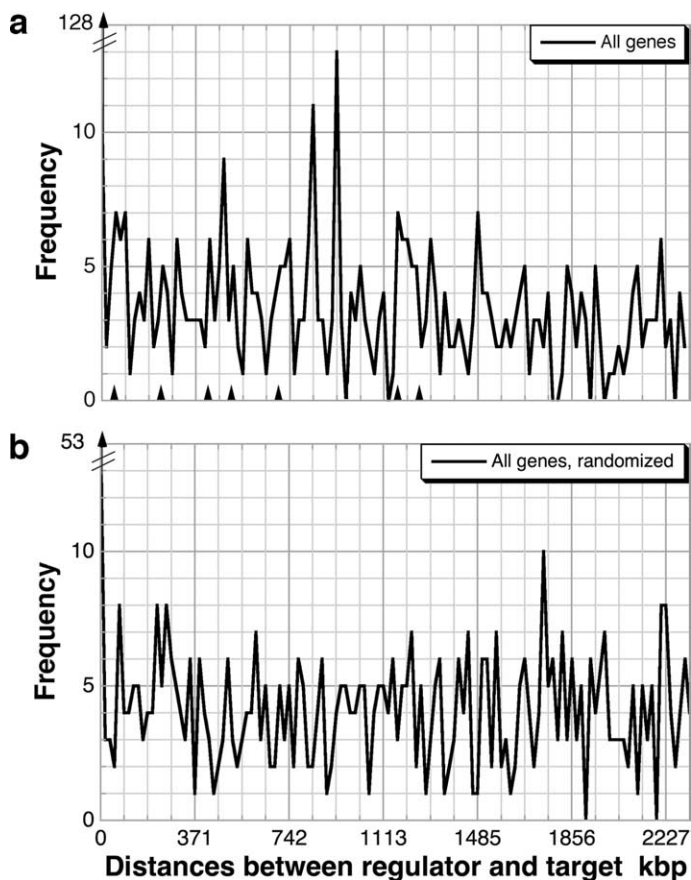


Figure 1. Distribution of distances separating a regulator-encoding cistron from its targets along the circular *E. coli* K12 MG1655 chromosome. Pairwise distances between translation startpoints were separately measured for each of the 116 regulators, using data from Shen-Orr et al.⁶ Only the first cistrons of target operons were considered. The resulting 555 pairwise distances are distributed together on these histograms with a grid interval of 92.8 kbp. The abscissa encompasses 2320 kbp, i.e. half the chromosome circumference. The distance between two successive plot points (or “bin size” for data discretization) has been varied without affecting the conclusions. A wide bin size of 18.56 kbp was chosen for this illustration, so that no smoothing was required. Calculations used Microsoft® Excel VBA routines. The routines and data are available upon request. a, Distances between all the regulator-encoding cistrons and their respective targets. The frequency at 0 kbp is 128. Arrowheads on the abscissa axis denote five peaks and two shoulders (frequency ≥ 5), positioned at mid-points between two gridlines. b,

Distances between all the regulator-encoding cistrons and their respective targets, after gene positions have been randomly attributed. Gene content, chromosome length and target lists are as in the natural chromosome. The frequency at 0 kbp is 53, due to self-regulations.

peaks on plots such as Figure 1a. Setting the threshold at 5, peaks fall on 18 of the 25 gridlines spaced by 92.8 kbp. Seven generally lower peaks fall at midpoints (arrowheads on the distance axis), suggesting the coexistence of a minor period of 46.4 kbp. No peak falls elsewhere (Figure 1a). For comparison, random models were built and compared to the data. For each random model, gene positions were attributed at random along a fake chromosome having identical length and gene content, while target lists remained unchanged. Such random models have no preferred period (e.g. upper curve in Supplementary Figure 1 online). Their distance distributions do not show periodicity and are less dispersed than real data (e.g. Figure 1b).

Since the raw data exposed on Figure 1 are noisy due to low numbers, they were subjected to a global, coarse-grained analysis on Figure 2. If the 25 successive segments of length 92.8 kbp in Figure 1a were superimposed on top of each other and set to unit length, the periodic peaks would now cumulate at a relative distance of ~ 0 , corresponding to the gridlines of Figure 1a. Crp is the only sequence-specific regulator in *E. coli* whose gene targets are in sufficient number^{5,6} to warrant a separate analysis (Figure 2a). All the other regulators were considered together (Figure 2b). In both cases, these regulator/target relative distances are more frequent around 0. By contrast, fluctuations are under 5% when gene positions are attributed at random as in Figure 1b (Figure 2c). For evaluation of the randomized set, it is essential to subtract self-regulations (gray), as they yield short distances whatever the gene position. Thus, it appears that Crp and most, if not all, other transcriptional regulators are encoded by genes that are positioned at regular intervals from their target genes.

So far, analysis made use of the genome sequence of the prototypical laboratory strain *E. coli* K12 MG1655. To check whether other *E. coli* chromosomes showed similar regularities, three additional strains were compared, including another K12 strain and two pathogenic strains whose genomes exhibit complex segmented

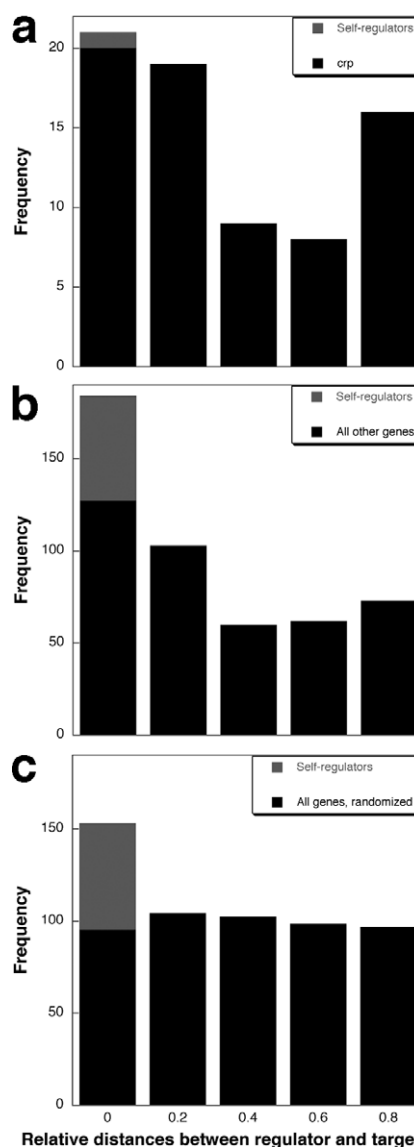


Figure 2. Relative distances separating a regulator-encoding cistron from its targets in *E. coli* K12 MG1655. The pairwise distances were measured as in Figure 1. The remainders of their division by the observed period (92.8 kbp) are distributed on this bar graph. Thus, frequencies falling on the gridlines of Figure 1 cumulate here at a relative distance of 0. Bar width is 18.56 kbp. Self-regulations are represented in gray. a, Relative distances between the Crp-encoding cistron and its targets. b, Relative distances between the cistrons encoding all the regulators, except Crp, and their respective targets. c, Relative distances between all the regulator-encoding cistrons and their respective targets, after gene positions have been randomly attributed.

Table 1. Comparison between four *E. coli* strains

Strain	Length (bp)	Orthologs	Relations	Self-regulations
K12 MG1655	4,639,221	–	555	58
K12 W3110	4,641,433	3974	541	57
CFT073	5,231,428	1765	79	15
O157:H7	5,528,445	1645	66	16
EDL933				

For each strain listed in the left column, are shown the chromosome length, the number of orthologous genes relative to K12 MG1655, the number of regulator/target relations⁶ that apply to these orthologs, and the number of self-regulations among these relations.

relationships to K12 (Table 1). Regulator/target relations⁶ were applied to the orthologous genes determined in each genome with respect to K12 MG1655. The low number of orthologs in the pathogenic strains (Table 1) did not allow a direct analysis of distance distributions. It was still possible to analyze relative distances for all available

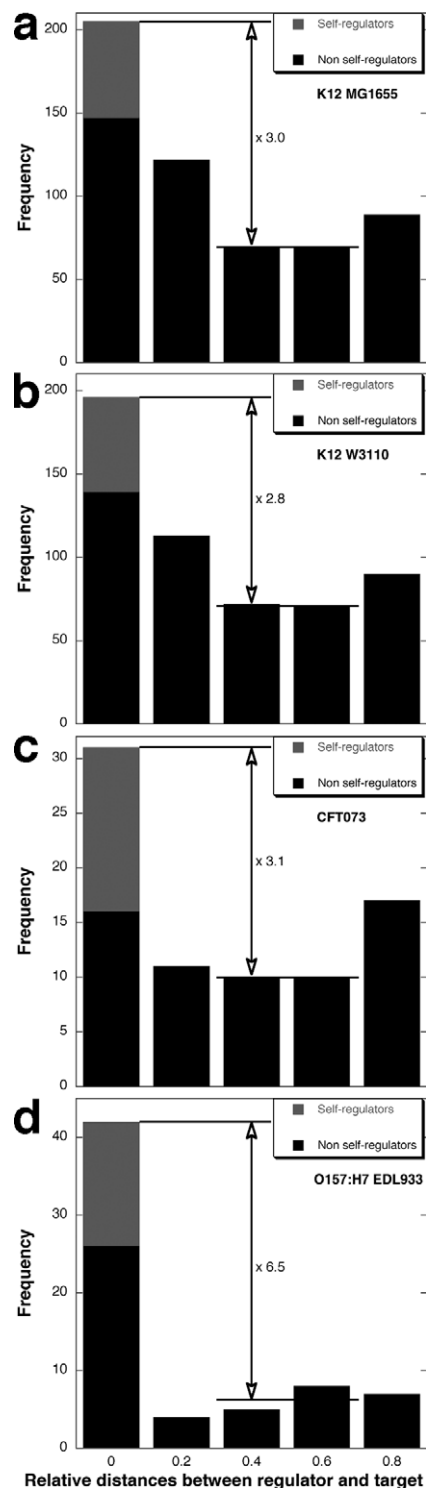


Figure 3. Relative distances separating a regulator-encoding cistron from its targets in four different *E. coli* strains. The distributions for all regulators were computed as in Figure 2. The ratio of highest (at abscissa 0) to lowest (average for abscissae 0.4 and 0.6) values is shown near the vertical arrow. The regulator/target relations were from Shen-Orr *et al.*⁶ The orthologs of the *E. coli* K12 MG1655 genes in other strains were retrieved in Feb 2004 from the Comprehensive Microbial Resource³² web pages of The Institute for Genomic Research at <http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl> or from Genbase FTP service at <http://ecoli.aist-nara.ac.jp/>.

regulators together (Figure 3). Figure 3a shows the distribution for K12 MG1655 and is therefore equivalent to the sum of the distributions shown in panels a and b of Figure 2 (all regulators, including Crp). The next panels of Figure 3 show the distributions for other strains, assuming 50 periods per chromosome. The ratio of frequency at relative distance 0 over mean frequency at distances 0.4 and 0.6 is indicated near the vertical arrow in each panel, and ranges among strains from about 3 to 6 (Figure 3). Thus, the genomes of these *E. coli* strains possess the same number of periods. Consequently, the periods are proportional to the overall chromosome length, which varies among strains by up to 20% (Table 1).

Periodicity of distances between coregulated targets

As the above periodicities constrain target positioning, secondary regularities were sought in the distances between pairs of targets. The distribution of such distances is shown in Figure 4 for the targets of Crp.^{5,6} It has a peak for the shortest distances, indicating 1-D target/target clustering.

However, for longer distances, the Crp targets tend to be regularly spaced by ~ 92.8 kbp or multiples thereof (Figure 4a; grid step is 92.8 kbp). There are very few “aperiodic” peaks that do not coincide with the gridlines, except in the last five intervals. However, these five aperiodic peaks, denoted by arrowheads on the distance axis, are spaced by ~ 92.8 kbp too (Figure 4a). When gene positions are attributed at random along a fake chromosome having identical length, gene content, and target lists, no periodicity and reduced dispersion are observed (Figure 4b).

For global analysis, the 25 successive 92.8 kbp segments in Figure 4a were superimposed as in Figure 2 (Figure 5a). Again, the relative distances between Crp coregulated genes are more frequent around 0, consistent with 3-D clustering. This trend is not observed when Crp target positions are attributed at random (Figure 5b). No significant regularity is detected for other regulators (e.g. Figure 5c), probably due to their low target numbers.

Regulator/target versus target/target patterns

The impact of the periodic phenomenon on target/target or regulator/target distances may crudely be estimated by two criteria. (i) Ratio of frequency at relative distance 0, over mean frequency at distances 0.4 and 0.6: for target/target distances, it is 1.4 for Crp, and ~ 1 for other regulators (Figure 5), while for regulator/target

The *E. coli* strains are: (a) K12 MG1655;³³ (b) K12 W3110;³⁴ (c) uropathogenic CFT073;³⁵ (d) enterohaemorrhagic O157:H7 EDL933.³⁶

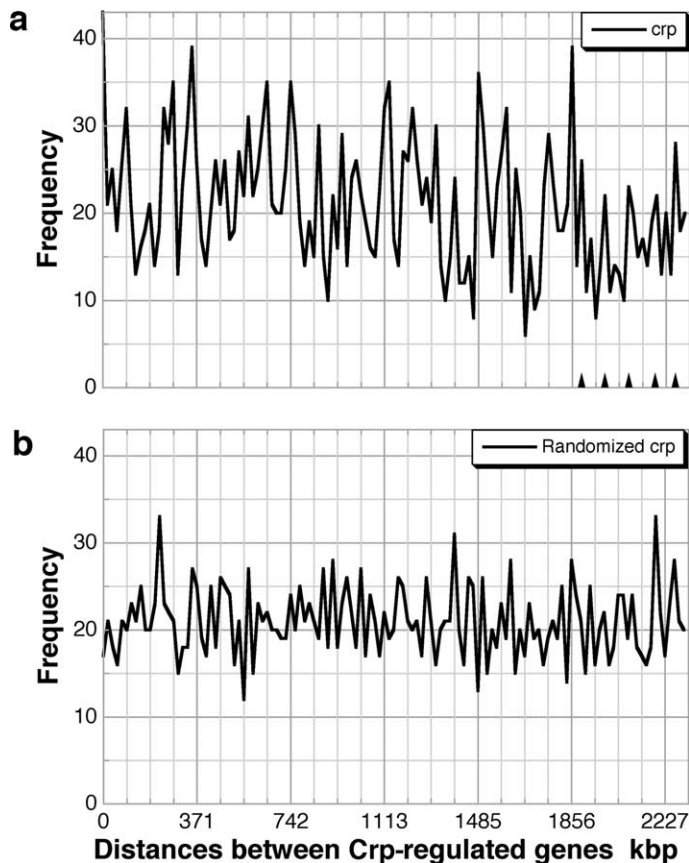


Figure 4. Distribution of distances separating Crp-regulated targets in *E. coli* K12 MG1655. The pairwise distances between the translation startpoints of all 2628 gene pairs in the list of Crp target genes⁶ are distributed on this histogram with a grid interval of 92.8 kbp. a, Distances between Crp-regulated targets. Arrowheads on the abscissa axis denote successive peaks above 20 that do not fall on the gridlines (they are still spaced by 92.8 kbp). b, Distances between Crp-regulated targets, after gene positions have been randomly attributed.

distances, it ranges from 3 to 6 (Figure 3). As midpoint peaks (Figure 1a) fall by definition in the 0.4–0.6 range, the latter ratios are underestimated. (ii) Deviation from randomness estimated by the Kolmogorov–Smirnov test: it is much less significant for Crp target/target than for all regulator/target distances (see Supplementary Table 1 online). These lower impacts indicate that the target/target pattern does not cause the regulator/target pattern. It suggests that the driving force of these periodicities is the proximity of the site of regulator production to its multiple DNA targets. In this view, the regular pattern of Crp coregulated targets is a mere consequence of these multiple positional relationships, consistent with the finding of an identical period for both phenomena.

Molecular and evolutionary basis

These regularities are consistent with the hypothesis of 3-D clustering. Clearly, regular spacing for one regulator does not demonstrate 3-D spatial proximity. However, the finding of an identical period for most regulators strongly constrains the interpretation, because their cistrons and their targets are interspersed throughout the whole chromosome. Only a generalized loop model, with a solenoidal rather than a radial or toroidal topology, can parsimoniously account for period invariance both for interspersed genes and for the whole genome. In other words, period invariance

must reflect the propagation along the whole chromosome of a local constraint generated by gene interspersing. As many regulators function simultaneously and some share targets, this local constraint presumably provides a potent self-organizational principle for the whole chromosome. Because the bacterial chromosome is circular, the regular spacing imposes that the period be an exact fraction of its full length. It appears to be the case, as the major period of 92.8 kbp is 1/50th of the K12 MG1655 chromosome circumference. Of course, the solenoidal topology must not be understood as a solenoidal geometry but as the most parsimonious means to satisfy a large set of spatial constraints that dynamically compete with others.

This peculiar epi-organization of the genome would not have emerged without an optimizing principle that selection can act upon to partially counteract the constant DNA flux in and out of the genome. Such a principle is proposed below, that unifies under the same mechanistic explanation both 1-D and 3-D clustering, while acknowledging that 1-D clustering has the added advantage of favoring coinheritance of functionally linked genes. If bacterial DNA were modeled as a compressed solenoid with one period per turn, the observed periodicities would optimize transcription by locally increasing the concentration of transcription factors and of their binding sites. They would moreover reduce the response delay by allowing the newly synthesized regulator to immediately

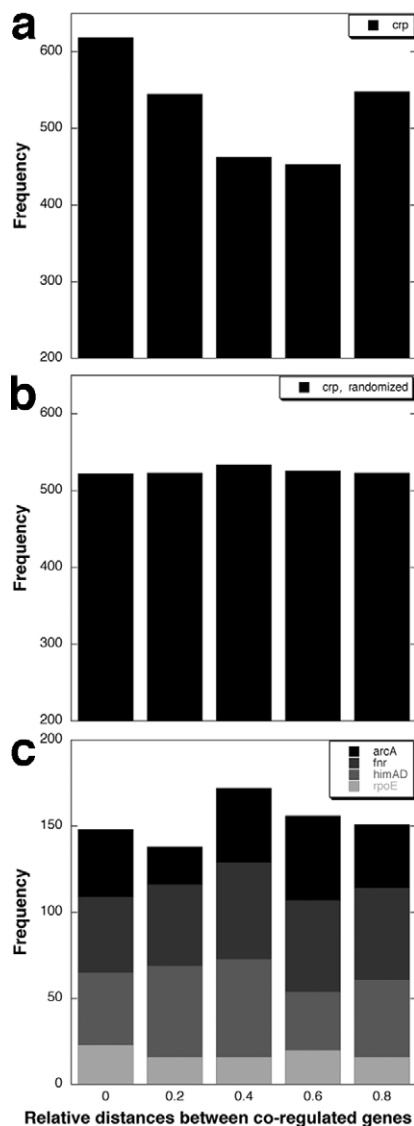


Figure 5. Relative distances separating co-regulated targets in *E. coli* K12 MG1655. The pairwise distances were measured as in Figure 4. The remainders of their division by 92.8 are distributed on this bar graph. Bar width is 18.56 kbp. a, Relative distances between Crp-regulated targets. b, Relative distances between Crp-regulated targets, after gene positions have been randomly attributed. c, Relative distances between genes regulated by ArcA, Fnr, HimAD (IHF), or RpoE.⁶

capture its right site without repeatedly binding/unbinding non-specific sites. This model is parsimonious because it simply extends the scope of local concentration effects from some well-documented 1-D intragenic cases¹ to a ubiquitous and often intergenic 3-D scheme. The specification of target and regulator sites can thus be viewed as a response to selective pressures that transcend those of operon-type organization. As far as regulator/target relations are concerned, this proposition is conceivable only for prokaryotes, where cotranscriptional translation³ confines regulator

production. By contrast, in eukaryotic yeast, 3-D clustering has been documented only for target/target relations.¹⁵

It is common practice to invert, insert or delete DNA segments in the *E. coli* chromosome. Generally, such genetic engineering is not obviously deleterious, although there are unexplained exceptions that often coincide with an abnormal nucleoid structure.^{16–18} More subtle effects are expected from an optimizing phenomenon, and they have not been looked for so far. Intriguing losses of sensitivity to regulators when promoters are moved from their chromosomal loci to plasmids may hint at such positional effects.¹⁹

Spatial clustering and chromosomal topology

The 3-D clustering model dictates that transcription foci and long-range DNA loops should dynamically self-organize in the nucleoid, especially around the most active transcription units. This prediction is consistent with the morphological demonstration of discrete foci^{7,8} and with the structural and biochemical observations of ~50 independent loop domains.^{9,20–25} However, transcriptional dynamics is central to the solenoidal interpretation of the observed periodicity. Indeed, starvation disperses the polymerases,⁸ presumably because reducing and spreading out the overall transcription activity weakens 3-D clusters. Also consistent with a dynamic view, *in vivo* structural data indicate that domain boundaries are transient¹² and that the number of loop domains drops as bacteria are grown on a poor medium¹¹ or enter stationary phase.¹² Another consequence of this dynamic view is that, within functional and physical limits, the number and length of the periods is an arbitrary number that should vary among species. This variability may explain the absence of regularity for gene clusters when 89 bacterial genomes are considered together.²⁶

Conclusion

Irrespective of their interpretation, the above results establish a close link between functional nucleoid structure and mRNA transcriptional regulation. It is noteworthy that this close link is inherent in the dual status of topoisomerases and “histone-like” structural proteins: these proteins modulate both DNA topology and gene expression in a growth phase-dependent manner.^{27–31}

Acknowledgements

I am grateful to Bernard Prum and Philippe Bouloc for critically reading this paper, and to members of the laboratory for help with the Kolmogorov–Smirnov test. Supported by funding

from CNRS, genopole® and Conseil Régional d'Ile-de-France.

References

- Müller-Hill, B. (1998). The function of auxiliary operators. *Mol. Microbiol.* **29**, 13–18.
- Dröge, P. & Müller-Hill, B. (2001). High local protein concentrations at promoters: strategies in prokaryotic and eukaryotic cells. *Bioessays*, **23**, 179–183.
- Miller, O. L., Jr, Hamkalo, B. A. & Thomas, C. A., Jr (1970). Visualisation of bacterial genes in action. *Science*, **169**, 392–395.
- Thieffry, D., Huerta, A. M., Pérez-Rueda, A. & Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis to transcriptional regulation in *Escherichia coli*. *Bioessays*, **20**, 433–440.
- Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Diaz-Peredo, E., Sanchez-Solano, F. *et al.* (2001). RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucl. Acids Res.* **29**, 72–74.
- Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31**, 64–68.
- Talukder, A. A., Hiraga, S. & Ishihama, A. (2000). Two types of localization of the DNA-binding proteins within the *Escherichia coli* nucleoid. *Genes Cells*, **5**, 613–626.
- Lewis, P. J., Thaker, S. D. & Errington, J. (2000). Compartmentalization of transcription and translation in *Bacillus subtilis*. *EMBO J.* **19**, 710–718.
- Pettijohn, D. E. (1996). *The Nucleoid* (Neidhardt, F. C., ed.), 2nd edit., vol. 1, pp. 158–166, American Society of Microbiology, Washington, DC.
- Woldringh, C. L. (2002). The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol. Microbiol.* **45**, 17–29.
- Sinden, R. R. & Pettijohn, D. E. (1981). Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc. Natl Acad. Sci. USA*, **78**, 224–228.
- Staccek, P. & Higgins, N. P. (1998). Gyrase and Topo IV modulate chromosome domain size *in vivo*. *Mol. Microbiol.* **29**, 1435–1448.
- Rosenfeld, N., Elowitz, M. B. & Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.* **323**, 785–793.
- Guelzim, N., Bottani, S., Bourguin, P. & Képès, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genet.* **31**, 60–63.
- Képès, F. (2003). Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J. Mol. Biol.* **329**, 859–865.
- Francois, V., Louarn, J., Patte, J., Rebollo, J. E. & Louarn, J.-M. (1990). Constraints in chromosomal inversions in *Escherichia coli* are not explained by replication pausing at inverted terminator-like sequences. *Mol. Microbiol.* **4**, 537–542.
- Niki, H., Yamaichi, Y. & Hiraga, S. (2000). Dynamic organization of chromosomal DNA in *Escherichia coli*. *Genes Dev.* **14**, 212–223.
- Guijo, M. I., Patte, J., del Mar Campos, M., Louarn, J.-M. & Rebollo, J. E. (2001). Localized remodeling of the *Escherichia coli* chromosome: the patchwork of segments refractory and tolerant to inversion near the replication terminus. *Genetics*, **157**, 1413–1423.
- Martin, R. G. & Rosner, J. L. (2002). Genomics of the marA/soxS/rob regulon of *Escherichia coli*: identification of directly activated promoters by application of molecular genetics and informatics to microarray data. *Mol. Microbiol.* **44**, 1611–1624.
- Worcel, A. & Burgi, E. (1972). On the structure of the folded chromosome of *Escherichia coli*. *J. Mol. Biol.* **71**, 127–147.
- Meyer, M., De Jong, M. A., Woldringh, C. L. & Nanninga, N. (1976). Factors affecting the release of folded chromosomes from *Escherichia coli*. *Eur. J. Biochem.* **63**, 469–475.
- Snyder, M. & Drlica, K. (1979). DNA gyrase on the bacterial chromosome: DNA cleavage induced by oxolinic acid. *J. Mol. Biol.* **131**, 287–302.
- Sinden, R. R., Carlson, J. O. & Pettijohn, D. E. (1980). Torsional tension in the DNA double helix measured with trimethylpsoralen in living *E. coli* cells: analogous measurements in insect and human cells. *Cell*, **21**, 773–783.
- Murphy, L. D. & Zimmerman, S. B. (2000). Multiple restraints to the unfolding of spermidine nucleoids from *Escherichia coli*. *J. Struct. Biol.* **132**, 46–62.
- Brunetti, R., Prosseda, G., Beghetto, E., Colonna, B. & Micheli, B. (2001). The looped domain organization of the nucleoid in histone-like protein defective *Escherichia coli* strains. *Biochimie*, **83**, 873–882.
- Audit, B. & Ouzounis, C. A. (2003). From genes to genomes: universal scale-invariant properties of microbial chromosome organisation. *J. Mol. Biol.* **332**, 617–633.
- Rouvière-Yaniv, J., Yaniv, M. & Germond, J. E. (1979). *E. coli* DNA binding protein HU forms nucleosome-like structure with circular double-stranded DNA. *Cell*, **17**, 265–274.
- Murphy, L. D. & Zimmerman, S. B. (1997). Stabilization of compact spermidine nucleoids from *Escherichia coli* under crowded conditions: implications for *in vivo* nucleoid structure. *J. Struct. Biol.* **119**, 336–346.
- Murtin, C., Engelhorn, M., Geiselmann, J. & Boccard, F. (1998). A quantitative UV laser footprinting analysis of the interaction of IHF with specific binding sites: re-evaluation of the effective concentration of IHF in the cell. *J. Mol. Biol.* **284**, 949–961.
- Schneider, R., Lurz, R., Lüder, G., Tolksdorf, C., Travers, A. & Muskhelishvili, G. (2001). An architectural role of the *Escherichia coli* chromatin protein FIS in organising DNA. *Nucl. Acids Res.* **29**, 5107–5114.
- Dame, R. T. & Goosen, N. (2002). HU: promoting or counteracting DNA compaction? *FEBS Letters*, **529**, 151–156.
- Peterson, J. D., Umayam, L. A., Dickinson, T. M., Hickey, E. K. & White, O. (2001). The comprehensive microbial resource. *Nucl. Acids Res.* **29**, 123–125.
- Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
- Itoh, T., Okayama, T., Hashimoto, H., Takeda, J., Davis, R. W., Mori, H. & Gojobori, T. (1999). A low rate of nucleotide changes in *Escherichia coli* K-12 estimated from a comparison of the genome sequences between two different substrains. *FEBS Letters*, **450**, 72–76.

35. Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D. *et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
36. Perna, N. T., Plunkett, G., Burland, V., Mau, B., Glasner, J. D., Rose, D. J. *et al.* (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, **409**, 529–533.

(Received 29 December 2003; received in revised form 22 April 2004; accepted 7 May 2004)

SCIENCE @ DIRECT®
www.sciencedirect.com

Edited by M. Yaniv

Supplementary Material comprising one Table and one Figure is available on Science Direct